

## Comparative Methods for the Analysis of Gene-Expression Evolution: An Example Using Yeast Functional Genomic Data

Todd H. Oakley,\*<sup>†1</sup> Zhenglong Gu,\* Ehab Abouheif,<sup>†2</sup> Nipam H. Patel,<sup>†2</sup> and Wen-Hsiung Li\*

\*Ecology and Evolution, and <sup>†</sup>Howard Hughes Medical Institute, University of Chicago

Understanding the evolution of gene function is a primary challenge of modern evolutionary biology. Despite an expanding database from genomic and developmental studies, we are lacking quantitative methods for analyzing the evolution of some important measures of gene function, such as gene-expression patterns. Here, we introduce phylogenetic comparative methods to compare different models of gene-expression evolution in a maximum-likelihood framework. We find that expression of duplicated genes has evolved according to a nonphylogenetic model, where closely related genes are no more likely than more distantly related genes to share common expression patterns. These results are consistent with previous studies that found rapid evolution of gene expression during the history of yeast. The comparative methods presented here are general enough to test a wide range of evolutionary hypotheses using genomic-scale data from any organism.

### Introduction

Since the Evolutionary Synthesis (Mayr and Provine 1980), which established genes as the primary source of evolutionary variation, a major focus of biology has been to understand the evolution of gene function. Despite this focus, some important measures of gene function are rarely used in evolutionary analyses. The evolution of gene function is perhaps most often studied by examining variation in the primary sequences of coding regions, an approach that emphasizes the biochemical and structural functions of genes. However, a gene's function can also be defined by its role in genetic pathways, cells, organs, tissues, organisms, and ecosystems. In fact, some have argued that studying spatial and temporal patterns of gene expression may be of paramount importance for understanding the genetic basis of evolutionary change (e.g., Britten and Davidson 1969; King and Wilson 1975; Wray et al. 2003). Gene-expression data are now increasing rapidly, despite being historically difficult to obtain quickly. For example, gene-expression data can be collected systematically on a large scale through microarray, expressed sequence tag (EST), and serial analysis of gene expression (SAGE). Furthermore, gene-expression patterns are being characterized for numerous genes in many different organisms by practitioners of the field of evolution of development (evo-devo). Despite this increasing database, gene-expression evolution is rarely modeled explicitly.

One approach to studying gene-expression evolution that does not rely on explicit evolutionary models is the comparison of expression profiles for pairs of duplicate genes (Wagner 2000; Gu et al. 2002b; Makova and Li 2003). With this approach, genetic distances are used as proxies for evolutionary time, inferred by comparing the coding regions of duplicate gene pairs. One prediction that

can be tested with the gene-pair approach is that genes separated by larger genetic distances (and, by assumption, more evolutionary time) should show larger differences in expression. Wagner (2000) concluded that this prediction was not met, as he failed to find a correlation between genetic distance and microarray expression divergence in yeast genes. Gu et al. (2002b) clarified this result, finding that such a correlation exists in closely related gene pairs, suggesting that gene expression evolves rapidly. In other words, shortly after duplication, genes tend to have similar expression patterns, but those patterns rapidly become distinct from each other. Using similar methodology, Makova and Li (2003) compared spatial expression of duplicated human genes, also concluding that gene-expression evolution is rapid, although still correlated with sequence divergence for a period of time after duplication.

Here, we introduce an explicitly model-based approach to investigating what processes shape gene-expression evolution. We employ general, diffusion-based maximum-likelihood models first described for the evolution of species' phenotypes (Felsenstein 1973; Mooers and Schluter 1998; Mooers, Vamosi, and Schluter 1999). We treat mRNA expression levels as traits of genes that have evolved during the history of yeast gene families, allowing us to map expression data on gene phylogenies estimated from sequence data. Like previous studies (Wagner 2000; Gu et al. 2002b), our expression data come from multiple microarray experiments, each of which quantified the genomic response to a perturbation, such as heat shocking, which we consider to be potential gene functions. Our evolutionary models assume that the expected variation in gene expression increases monotonically with some measure of the time available for change. The time available for change may be estimated from the genes' phylogenies in different ways, constituting different models of evolution (Mooers and Schluter 1998; Mooers, Vamosi, and Schluter 1999). For example, we can use estimated genetic distance or the number of duplication events between two genes to estimate time available for change in expression. This approach allows us to consider previous ideas about gene-function evolution, such as neofunctionalization (the acquisition of a new function by one copy of a duplicate gene) or subfunctionalization (the partitioning of functions between duplicate

<sup>1</sup> Present address: Ecology Evolution and Marine Biology, University of California-Santa Barbara.

<sup>2</sup> Present address: Department of Integrative Biology, University of California, Berkeley.

Key words: evolutionary biology, gene expression, nonphylogenetic, genomic scale.

E-mail: oakley@lifesci.ucsb.edu.

*Mol. Biol. Evol.* 21(12):1–11. 2004

doi:10.1093/molbev/msh257

Advance Access publication Month X, XXXX

gene pairs). Our models use a maximum-likelihood framework, so we can use standard statistical methods to compare nine different models in three general classes, each with different implications for the evolution of gene expression. The example presented show that nonphylogenetic models best fit yeast gene-expression data, a result consistent with previous studies. The methods can be used generally with different data sets to increase understanding of processes that underlie the evolution of gene expression or other measures of gene function, such as fitness effects of gene deletion. While the current manuscript was in review, Gu (2004) independently proposed a similar framework for the investigation of gene-expression profiles. Throughout the rest of this paper, we compare the models and results of his approach to our own.

## Methods

Our methods can be divided into five sequential steps. First, we partition all yeast genes into families. Second, we perform phylogenetic analyses separately on each of the 10 largest gene families, resulting in a “gene tree” for each family. Third, we collect microarray expression data from the literature for all genes. Fourth, we establish nine maximum-likelihood models for gene-expression evolution that relate the levels of expression to gene trees. Fifth, we compare likelihood values of different models to test specific hypotheses about gene-expression evolution. Below, we describe each of these steps in greater detail. Notice the important point that gene trees are not estimated from the expression data, but from the gene-sequence data. These are separate and independent data sets, so no circularity exists.

### Partitioning the Yeast Proteome

All yeast genes were partitioned into families, as described previously (Gu et al. 2002a). We compared each protein in turn to each other protein using FASTA. To be grouped into a single family, two proteins must meet two criteria. First, they must contain a FASTA-alignable ( $E = 10$ ) region that comprises greater than 80% of the longer protein. Second, two genes must be more similar to each other than to a specified threshold, described previously (Rost 1999; Gu et al. 2002a). We chose the 10 largest gene families (in terms of number of genes) for further analysis. We chose a stringent method for grouping genes into families, and it is likely that we may not have identified distantly related genes in some gene families. This stringency will be conservative with respect to our conclusions. Because we report support for nonphylogenetic methods consistent with rapid evolution of gene expression, including more distantly related genes would only increase such support.

### Phylogenetic Analyses

Next, we separately performed phylogenetic analyses on each gene family. We aligned amino acid sequences using default parameters of ClustalW (Higgins, Bleasby, and Fuchs 1992), as implemented in BioEdit ([http://](http://www.mbio.ncsu.edu/BioEdit/bioedit.html)

[www.mbio.ncsu.edu/BioEdit/bioedit.html](http://www.mbio.ncsu.edu/BioEdit/bioedit.html)), and then back translated to the original (but now aligned) DNA sequences.

We used maximum-likelihood (ML) analysis, estimating phylogenetic trees for 10 yeast gene families using PAUP\* version 4.0b10 (Swofford 1999) and the Tamura-Nei (Tamura and Nei 1993)+ Gamma + Invariant Sites model of evolution. We selected this model because it was the best-fit model for the largest gene family, as determined by likelihood ratio tests in ModelTest (Posada and Crandall 1998). To facilitate comparison among different gene families, we used the same model for all analyses, even though it may not be best fit for every gene family analysis. Because we had no information on outgroups for the gene families in question, we used midpoint rooting. Another possible approach was taken by Gu (2004), who determined outgroup status by analyzing yeast gene families in the context of orthologous genes from other species. We assumed a molecular clock in the phylogenetic analyses to allow the estimation of lengths for all branches in a rooted phylogeny (nonclock methods assign a length 0 to one branch at the root of the tree). Here, we report analyses that assumed a molecular clock, however additional analyses (not shown) using nonclock methods for branch-length estimation did not qualitatively change the final results.

### Gene-Expression Data

Microarray expression data exist for almost the entire yeast proteome. Experimentalists have perturbed yeast in different ways, for example by heat shocking, to quantify changes in gene expression caused by the experimental perturbation (DeRisi, Iyer, and Brown 1997; Chu et al. 1998; Spellman et al. 1998; Gasch et al. 2000; Lyons et al. 2000). After these perturbations, changes in gene expression were usually measured at different timepoints past the perturbation.

We used cDNA microarray data, which are presented as ratios of the initial level of gene expression at different timepoints past the perturbation to the initial level at time 0 (DeRisi, Iyer, and Brown 1997; Chu et al. 1998; Spellman et al. 1998; Gasch et al. 2000; Lyons et al. 2000). The ratios are then  $\log_2$  transformed, so that the final numbers represent number of doublings in gene expression with respect to the original timepoint. There is no attempt at quantifying the absolute level of expression. Therefore, for these data, our models predict that the variance of the number of doublings in gene-expression level increases in proportion to evolutionary time. Using a notation similar to Gu (2004), a typical data set for a given family of size  $M$  includes

$$\begin{bmatrix} \text{gene\_family} \\ i = \text{gene1} \\ i = \text{gene2} \\ \dots \\ i = \text{geneM} \end{bmatrix} = \begin{bmatrix} k = 1 & k = 2 & \dots & k = N \\ D11 & D12 & \dots & D1N \\ D21 & D22 & \dots & D2N \\ \dots & \dots & \dots & \dots \\ DM1 & DM2 & \dots & DMN \end{bmatrix},$$

where  $k$  represents multiple microarray experiments (the different perturbations such as heat shock or acid treatment). Each element in the matrix above ( $D11 \dots DMN$ )

is a vector of data that represent expression levels at multiple timepoints. For example,  $DII$  might equal  $\{r_1, r_2, \dots, r_t\}$ , where  $r_1$  through  $r_t$  are the  $\log_2$  transformed ratios of gene-expression level at time  $t$  to the gene-expression level at time 0. Note that different microarray experiments ( $k$  columns above) may differ in the number of timepoints measured, but an individual experiment usually has equal numbers of timepoints across different gene families ( $i$  rows above), unless data are missing because of experimental difficulties. Most available continuous character-likelihood methods cannot deal with such missing data (Felsenstein 1973). Therefore, within a given gene family, we removed timepoints that were missing data for one or more genes. This caused some experimental perturbations to be removed for some gene families. This also currently precluded the use of oligonucleotide microarray data, which were mostly incomplete.

An important assumption of continuous character-likelihood methods is that the traits are independently, identically distributed (i.i.d.) (Felsenstein 1973, 1988, 2004; Gu 2004). For yeast microarray data, as analyzed here, the i.i.d. assumption may be violated in two different ways (Gu 2004). First, there may be experimental correlations; for example, when expression levels at different timepoints are correlated. Second, i.i.d. may be violated by phylogenetic correlations, where evolutionary changes in expression are correlated in different genes, perhaps as a result of being members of the same genetic pathway.

Gu (2004) investigated the i.i.d. assumption, concluding that although the level of correlation was nontrivial, maximum-likelihood parameter estimates were similar whether or not he accounted for such correlation. Gu (2004), therefore, concluded that “likelihood under the i.i.d. assumption is useful for fast and large-scale analyses,” such as those we present here. Nevertheless, we attempted to reduce experimental correlation in one simple way: After we found that expression data from adjacent timepoints were correlated in our data, we instead used values of the change in expression from one timepoint to the next for further analysis, which showed less correlation.

Another important consideration for the data that we used is the level of cross-hybridization. cDNA data are likely to contain some signal that is caused by cross-hybridization, as RNA from one gene will likely hybridize to similar motifs in multiple different genes on a microarray. Cross-hybridization could affect our methods by artifactually increasing the estimated similarity of gene-expression patterns for more closely related genes. Cross-hybridization would, therefore, decrease the level of support for our nonphylogenetic models, which assume that closely related genes are no more likely to share gene-expression patterns than more distantly related genes. Cross-hybridization may also increase support for distance models, which predict that genetic distance is a good predictor of gene-expression similarity. We discuss cross-hybridization in light of our specific results later. Note that cross-hybridization is a concern for the primary data used for the example presented here because we used cDNA microarray data, but the methods presented are not limited to use with cDNA microarray data.

## Modeling Gene Expression

We used simple diffusion (Brownian motion) models, originally designed for analyses of species’ phenotypes evolving along a species’ phylogeny, to model the evolution of gene expression in gene families (Felsenstein 1973; Martins and Garland 1991; Mooers, Vamosi, and Schluter 1999). The merits, assumptions, limitations and advantages of using such models in evolution have been discussed at length elsewhere (Felsenstein 1988; Diaz-Uriarte and Garland 1996; Mooers and Schluter 1998). These models assume that the expected variation in phenotype (in this case gene expression) increases monotonically with some measure of time. For a particular set of phenotype values at ancestral nodes, the general case of the likelihood ( $L$ ) of observing a set of phenotypic data for a single character at the tips of a bifurcating rooted phylogenetic tree is given by:

$$L = \prod_{n=1}^{2N-1} \frac{1}{\sqrt{2\pi(v_{n1} + v_{n2})}} \exp \left[ -\frac{(x_{n1} - x_{n2})^2}{2(v_{n1} + v_{n2})} \right] \quad (1)$$

Here, the likelihood is the product over all nodes on the tree;  $n$  represents each node and  $N$  is the number of tips on the tree. The term  $(x_{n1} - x_{n2})$  is the differences in phenotype values at the two descendents of each node  $n$ . Finally,  $v_{n1}$  and  $v_{n2}$  are variance values. In practice, these are derived from the branch lengths of the phylogeny in units of expected amount of time available for phenotypic change along branches of the tree. To calculate  $L$ , Felsenstein (1981b) described the restricted-evolution maximum-likelihood (REML) algorithm, which integrates over all possible phenotypic states at each node. This integration is performed by making two adjustments to equation 1 (which only holds when nodes are tips of the phylogenetic tree in question). First, the phenotype values ( $x_{n1}$  and  $x_{n2}$ ) of internal nodes are set to equal weighted averages of their descendents (see equation 10 in Felsenstein [1985]). Second, the branch lengths ( $v_{n1}$  and  $v_{n2}$ ) for internal branches are augmented (see equation 10 in Felsenstein [1985]).

Equation 1 can also be written in a different way (Felsenstein 1973; Mooers, Vamosi, and Schluter 1999):

$$L = \prod_{i,i'} \frac{1}{\sqrt{2\pi\beta t_i'}} \exp \left[ -\frac{(x_i - x_{i'})^2}{2\beta t_i'} \right] \quad (2)$$

Here,  $x_i$  is the phenotype at node  $i$  and  $x_{i'}$  is the phenotype of a descendent node ( $i'$ ) of  $i$ .  $\beta$  is the rate of phenotypic change, and  $t_i'$  is the time available for evolutionary change between the nodes  $i$  and  $i'$  (along a single branch). The product is taken over all branches on the tree. Like equation 1, equation 2 is for a specific set of ancestral phenotypes. To calculate  $L$ , we integrate over all possible ancestral states using REML as described above for equation 1. Most importantly for the current discussion, our different models for gene-expression evolution are constructed by changing assumptions about the evolutionary time available for change along branches of gene trees ( $t_i'$ ).

We tested three specific types of model within each of three general classes, yielding a total of nine different

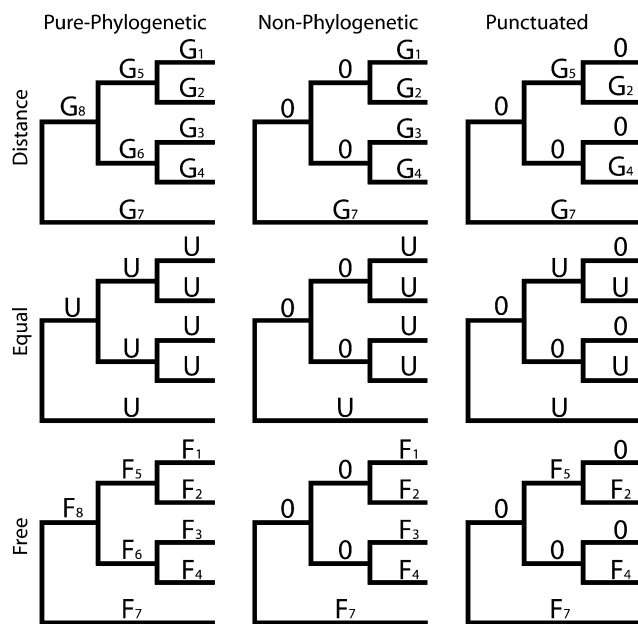


FIG. 1.—Nine different maximum-likelihood models of gene-expression evolution. These models predict that change in expression increases monotonically with the “time” available for change. Time available for change is estimated in different ways for different models, as indicated by different letters above branches of a hypothetical gene tree that would be estimated from sequences of a gene family. Branches labeled “G<sub>*i*</sub>” assume expression change is equal to genetic distance of that branch, those labeled “U” assume a unit (equal) amount of change, and those labeled “F<sub>*i*</sub>” are estimated from the expression data itself (free). Branches labeled “0” assume no change has occurred. Columns represent three different classes of models. The pure phylogenetic class assumes expression change occurs on every branch of the phylogeny, the nonphylogenetic class assumes expression change occurs only along terminal branches, and the punctuated class assumes expression change occurs on only one of every pair of descendent branches.

models. Model types vary in how the branch length values  $t_{i'}$  are specified in equation 2 (fig. 1), in essence, varying the amount of time that we assume is available for expression divergence. The first specific model type assumes that genetic distance predicts the amount of divergence in gene expression. In this case, the times available for change in gene expression ( $t_{i'}$  from equation 2) are set equal to the genetic distances that were calculated from gene-sequence data using phylogenetic methods. Note that we used overall measures of genetic distance based on entire coding regions of genes, but many other measures could also be used with different implications, such as genetic distance estimates derived from nonsynonymous sites, synonymous sites, 5' flanking regions, certain subregions of the protein, and radical nonsynonymous sites. Each of these different distance measures could be compared with each other to test which model is the best predictor of expression divergence.

The second model type consists of “equal models” (Mooers, Vamosi, and Schluter 1999). Here, the  $t_{i'}$  parameters are set equal for every branch. When all branches are assumed to have equal divergence in expression, more change in expression is expected to occur with more duplication events. For the third specific model, the  $t_{i'}$  parameter for each unconstrained branch is estimated

from the expression data itself by maximizing the likelihood function. This model does not assume a constant rate of expression change. Like Mooers, Vamosi, and Schluter (1999), we call these “free models.”

Each of the three specific model types (genetic distance, equal, and free) was implemented in three different general classes for a total of nine different models (fig. 1). The first general class is “pure phylogenetic” models, which assume that the branching patterns of the genes can predict expression. The pure phylogenetic class allows for change in gene expression along every branch of a gene phylogeny. For example, under the pure phylogenetic/equal model, all branches of a gene tree are fixed equal to each other. Second, the “nonphylogenetic” class of models assumes that related genes are no more likely than more distantly related genes to share similar expression patterns. The nonphylogenetic class of models assumes that no change occurred in gene expression in the internal branches of a phylogeny. All change is restricted to the terminal branches, equivalent to assuming there is no phylogenetic signal in the data (Mooers, Vamosi, and Schluter 1999). Finally, the “punctuated” model class assumes that the gene phylogeny can predict expression; however, at every branching point one daughter gene changes expression, whereas the other does not. Implementation of most of these models has been done previously (Mooers and Schluter 1998; Mooers, Vamosi, and Schluter 1999), with the exception of punctuated models. Below, we describe implementation of punctuated models.

To specify a model where change in only one descendent of each node is allowed, we implemented a specific case of the more general REML algorithm (equation 1). It is straightforward to restrict the allowable character change so that only one descendent of every node is allowed to change in phenotype. Under such a model, one and only one of each pair of  $v_{ni}$  values is set to 0 in equation 1. When the variance ( $v_{ni}$ ) equals 0, no change in phenotype is allowed along that branch of the tree (i.e., there is no variation).

Given this restriction, it follows that at each node of a tree, there are two possible combinations of variance values (i.e., branch lengths): ( $v_{n1} = \text{free}, v_{n2} = 0$ ) and ( $v_{n1} = 0, v_{n2} = \text{free}$ ). Let  $L_{n10}$  be the likelihood of observing the given phenotypes at the two descendents of node  $n$ , conditional upon  $v_{n1}$  being free to vary and  $v_{n2}$  being set equal to 0. Similarly, let  $L_{n01}$  be the likelihood conditional upon  $v_{n1}$  being equal to 0 and  $v_{n2}$  being free to vary. The number of different combinations of variance values for the nodes on an entire tree is then  $2^{N-1}$ , where  $N$  is the number of tips on the tree (two possibilities at each node and  $N-1$  total nodes in a rooted bifurcating tree). The likelihood of a given set of data assuming our punctuated mode of character change can be given by the equation

$$L = \prod_{n=1}^{2^N-1} (0.5L_{n01} + 0.5L_{n10}) \quad (3)$$

In equation (3), the conditional likelihood values based on the two allowable combinations of variance values at each node  $n$  are summed. Each of these terms is multiplied by a prior probability, assumed in this equation to be equal to

each other (0.5). The product is taken over all nodes of the phylogenetic tree. We can make equation (3) more explicit by replacing  $L_{n10}$  and  $L_{n01}$  with likelihood formulas derived from equation (1):

$$L = \prod_{n=1}^{2N-1} \left\{ 0.5 \frac{1}{\sqrt{[2\pi(v_{n1} + 0)]}} \exp \left[ -\frac{(x_{n1} - x_{n2})^2}{2(v_{n1} + 0)} \right] + 0.5 \frac{1}{\sqrt{[2\pi(0 + v_{n2})]}} \exp \left[ -\frac{(x_{n1} - x_{n2})^2}{2(0 + v_{n2})} \right] \right\} \quad (4)$$

The approach of summing over all combinations of  $L_{n10}$  and  $L_{n01}$ , suggested by P. Lewis of the University of Connecticut, is analogous to summing likelihood values over all possible ancestral states, which is done in likelihood analyses of DNA sequences (e.g., Felsenstein 1981a) and discrete morphological characters (Pagel 1994, 1999; Lewis 2001). This approach allows the comparison of likelihood values for punctuated models with other models, such as those described above. The number of degrees of freedom is then equal to the number of free variance parameters in equation (3), which is as many as  $N-1$  (one branch at every node may be free to vary). Note also that all of the free variance parameters can be constrained to be equal to each other (in this case there is only one free parameter), which is equivalent to assuming that when phenotypic change occurs, the magnitude is always similar.

Some of the models we present here are equivalent to those presented recently by Gu (2004), who presented four models termed the Brownian Motion (B), Lineage Specific (L), Directional Trend (D), and Dramatic Shift (S) models. Our phylogenetic/equal model is equivalent to Gu's B model, our phylogenetic/free model is equivalent to Gu's L model, and our phylogenetic/punctuated model is similar to Gu's S model. We do not present an equivalent of Gu's Directional Trend (D) model, and we note that our non-phylogenetic and genetic-distance models have no counterpart in Gu (2004).

To calculate likelihood values for each of the nine models, we created the computer program CoMET (Continuous-character Model Evaluation and Testing), which is available from T.H.O. CoMET uses code from the PHYLIP program CONTRAST (Felsenstein 1995) to calculate the likelihood values of each of our nine models in a manner that is independent of the scale of characters and branch lengths. Each of the three "free" models is already independent of scale; however, we also made the other six models independent of scale. To do so, the  $ti'$  parameter for every branch was multiplied by a scaling parameter, which itself took the value that maximizes the overall likelihood of a given model (Mooers and Schluter 1998; Mooers, Vamosi, and Schluter 1999). The results of CoMET are similar to those of the program FIT (Mooers, Vamosi, and Schluter 1999), except that CoMET calculates the likelihood of several additional models and can calculate automatically likelihood values for all models, numerous trees, and multiple data sets, allowing genomic-scale analyses to be performed. CoMET also uses rooted trees, whereas FIT uses unrooted trees. This difference makes some resulting likelihood values slightly different

between FIT and CoMET; for example, a rooted tree where all branch lengths are set equal has one more internal branch than an unrooted tree, which changes the resulting likelihood value. Note that punctuated models require a rooted tree.

#### Comparing Model Likelihood Values

In general, the models we present here can be compared using standard techniques in maximum likelihood (Edwards 1992). For the current study, we compared the likelihood values of different models in two different ways. First, we tested specific hypotheses using standard likelihood comparison procedures. Second, we used the Akaike Information Criterion (AIC), which allows the direct comparison of likelihood models with different numbers of parameters (Akaike 1973).

To illustrate our maximum-likelihood approach, we tested two hypotheses from the previous literature. First, we tested whether a general lack of fit exists between genetic distance and divergence in gene expression (Wagner 2000). Using our models, this leads to the prediction that the genetic-distance models fit worse than equal or free models. All of our distance and equal models have one parameter (rate of evolution); therefore, log-likelihood values may be compared directly, with a difference of 2 being considered significant (Mooers, Vamosi, and Schluter 1999). Free models have more parameters and may be compared with other models using standard likelihood ratio tests (Mooers, Vamosi, and Schluter 1999). Second, we tested the hypothesis that gene expression evolves rapidly, with more closely related genes sharing expression patterns for a short time (Gu et al. 2002b). Using our models, this hypothesis predicts that nonphylogenetic/distance models fit the data better than a pure phylogenetic/distance model. In other words, genetic distances since the most recent gene duplication events are better predictors of mRNA expression change than are overall genetic distances between two genes. Here again, these models both have one parameter and may be considered significantly different if their log-likelihood values differ by 2 or more (Mooers, Vamosi, and Schluter 1999).

We also used the AIC, a standard formula used to compare maximum-likelihood models with different numbers of parameters:

$$AIC = -2 \ln L + 2P \quad (5)$$

where the likelihood,  $L$ , is calculated by equation (1) and  $P$  is the number of parameters in the given model. Although not strictly a statistical significance test, the model with the lowest AIC value is considered best fit. For each data set, we ranked the nine different models using AIC. In our models,  $P = 1$  for all genetic distance and equal models because only  $\beta$  is estimated (Mooers, Vamosi, and Schluter 1999). Free models maximize each unconstrained branch and so if  $G =$  number of genes, the pure phylogenetic/free model uses  $P = 2G - 2$  (Mooers, Vamosi, and Schluter 1999), the nonphylogenetic/free model uses  $P = G$ , and the punctuated/free model uses  $P = G - 1$ .

**Table 1**  
**Yeast Gene Families Used in the Current Study**

Size	Possible Function	Gene List
18	Hexose transporters	YDL245C, YDR342C, YDR343C, YDR345C, YEL069C, YFL011W, YHR092C, YHR094C, YHR096C, YIL170W, YJL214W, YJL219W, YJR158W, YLR081W, YMR011W, YNL318C, YNR072W, YOL156W
17	Permeases	YBR068C, YBR069C, YBR132C, YCL025C, YDR046C, YDR508C, YEL063C, YFL055W, YGR191W, YKR039W, YLL061W, YNL268W, YNL270C, YOL020W, YOR348C, YPL265W, YPL274W
13	Helicases	YDR545W, YEL077C, YER190W, YGR296W, YIL177C, YJL225C, YLL066C, YLL067C, YLR466W, YLR467W, YML133C, YNL339C, YPL283C
11	DUP (unknown function)	YBR302C, YDL248W, YFL062W, YGL263W, YGR295C, YHL048W, YJR161C, YKL219W, YML132W, YNL336W, YNR075W
11	GTP-binding	YBR264C, YCR027C, YER031C, YFL005W, YFL038C, YGL210W, YKR014C, YLR262C, YML001W, YNL093W, YOR089C
10	Heat shock proteins	YAL005C, YBL075C, YDL229W, YEL030W, YER103W, YJL034W, YJR045C, YLL024C, YLR369W, YNL209W
8	ABC transporters	YDR011W, YDR406W, YIL013C, YNR070W, YOR011W, YOR153W, YOR328W, YPL058C
7	$\alpha$ -Glucosidases	YBR299W, YGR287C, YGR292W, YIL172C, YJL216C, YJL221C, YOL157C
7	ADP-ribosylation	YBR164C, YDL137W, YDL192W, YMR138W, YOR094W, YPL051W, YPL218W
7	Kinases	YBL016W, YBR160W, YDL108W, YGR040W, YMR139W, YPL031C, YPR054W

The AIC is deceptively simple and is based on explicit information theory mathematics. Several variants of the AIC also exist, as do similar model selection criteria (Royall 1997). Any of these can be used with the general likelihood framework that we present, depending on the specifics of the data at hand. We present here only the results of the AIC as one illustration of our methods.

#### An Example

To partition the yeast proteome into families, we used methods identical to Gu et al. (2002a), and detailed results are presented there. For the current study, we analyzed the 10 largest gene families as detailed in table 1. Next, we performed phylogenetic analyses on the gene families. Results for ML analyses are presented in figure 2, and parameter estimates are presented in the Supplementary Material online. Assuming these 10 phylogenetic trees, we calculated likelihood values for each of nine different

models of gene-expression evolution for 14 different perturbation experiments (Supplementary Material online contains all likelihood values).

Our comparisons of likelihood values supported two previous hypotheses. First, genetic distance models fit the data significantly worse than equal or free models: specifically the sum of the likelihoods for all comparisons (which are presented in Supplementary Material online) was orders of magnitude lower for genetic-distance models than for equal or free models. Second, the nonphylogenetic/distance model had a higher likelihood than a pure phylogenetic/distance model in 119 of 152 function/gene family analyses (highly significant in a binomial test), consistent with the hypothesis of rapid evolution of gene expression (Gu et al. 2002b; Makova and Li 2003).

We used AIC to rank each model within a specific function/gene family analysis. Figure 3 shows the results of this ranking procedure. Nonphylogenetic models had the best AIC score in 117 of 152 function/gene family

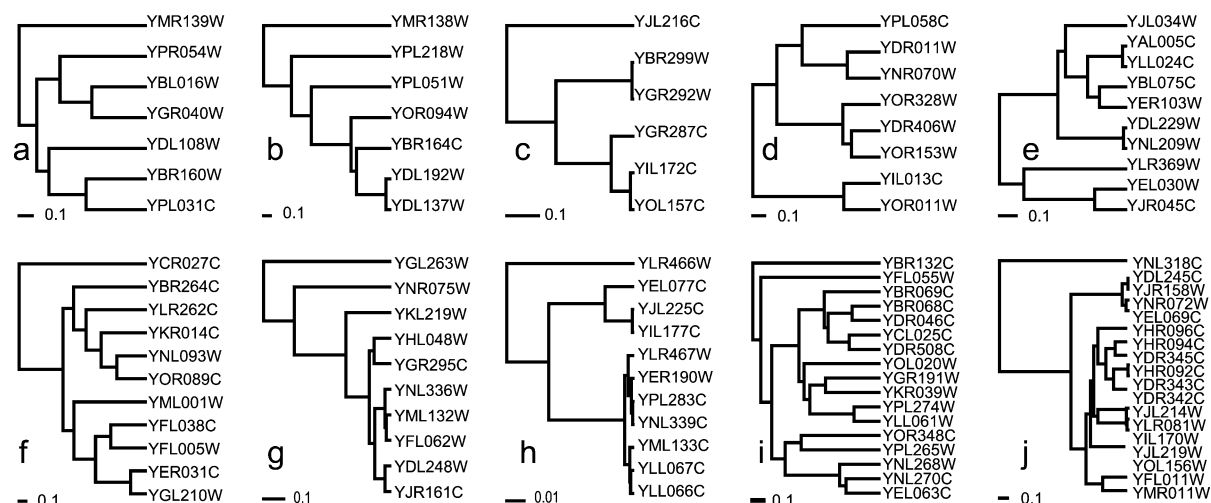


FIG. 2.—Results of phylogenetic analyses of the 10 largest yeast gene families. We used maximum likelihood, assuming a Tamura-Nei + Gamma/Invariant sites model and a molecular clock. (a) Kinases, (b) ADP-ribosylation, (c)  $\alpha$ -glucosidases, (d) ABC transporters, (e) heat shock proteins, (f) GTP-binding proteins, (g) “DUP” gene family, unknown function, (h) helicases, (i) permeases, and (j) hexose transporters. Relative branch lengths are proportional to number of substitutions per site and different trees are drawn to different scales.

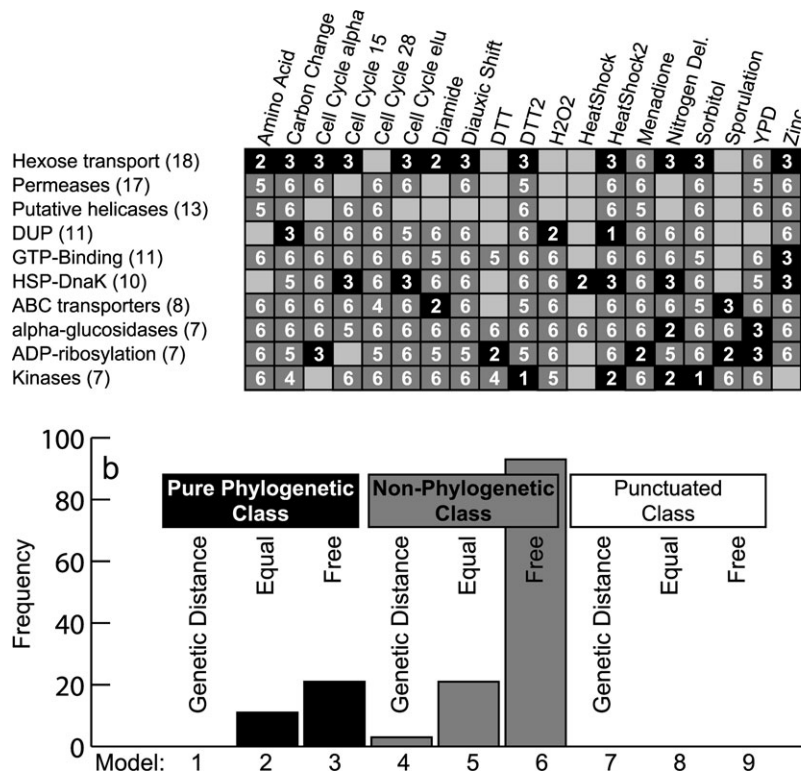


FIG. 3.—Results of Akaike Information Criterion (AIC) test. (a) The model with the best AIC value is indicated for each gene family/“function” combination. Models numbers are 1 = pure phylogenetic/genetic distance; 2 = pure phylogenetic/equal; 3 = pure phylogenetic/free; 4 = nonphylogenetic/genetic distance; 5 = nonphylogenetic/equal; 6 = nonphylogenetic/free; 7 = punctuated/genetic distance; 8 = punctuated/equal; 9 = punctuated/free. (b) Histogram showing number of times each model was favored by AIC for all gene family/“function” pairs.

analyses. In addition, genetic distance models had the lowest number of number one ranks within each model class (e.g., nonphylogenetic and phylogenetic), again consistent with a lack of general correlation between genetic distance and change in expression level.

## Discussion

We developed nine different models for gene-expression evolution, based on Brownian motion maximum-likelihood methods, which have been used previously to describe the evolution of species’ phenotypes. The nine models presented here differ from each other in how the evolutionary time available for change in gene expression is measured, allowing us to determine which of these estimates of evolutionary time best describe observed differences in gene expression. Interestingly, these different models correspond well with previous ideas about gene-function evolution, which largely come from studies of gene duplication. Our methods utilize the phylogenetic information present in gene families, and every node of each gene family tree represents a gene-duplication event, allowing our models test well-known ideas about gene duplication evolution (table 2).

The most widely supported models in the current study were the nonphylogenetic class of models. Under nonphylogenetic models, more closely related genes are no more likely than more distantly related genes to share similar expression patterns. Despite the fact that cross-

hybridization that is likely present in cDNA microarray data would lower support for nonphylogenetic models (discussed in *Methods*), nonphylogenetic models were the best-supported models. Another bias with the opposite consequences may come into play because data that lack evolutionary signal altogether also would favor nonphylogenetic models. This may lead to overestimated support for nonphylogenetic models because any given gene family probably lacks involvement in many physiological functions. Therefore, noise rather than evolutionary signal would dominate expression data for those function/gene family combinations. For example, if helicase genes show no response to heat shock, we would expect no evolutionary signal in that analysis and support for nonphylogenetic models would be overestimated. Although we recognize these potential biases, we favor the interpretation that the observed support of nonphylogenetic models is caused by the erasure of historical signal during the rapid evolution of gene expression. The rapid evolution of gene expression is a result supported by other empirical studies of yeast gene-expression data (Ferea et al. 1999; Gu et al. 2002b).

A second main result is the poor fit of genetic-distance models in comparison with equal or free models. This result can be understood in light of previous studies on pairs of duplicate genes that paid specific attention to the correlation of genetic distances and gene-expression divergences (Wagner 2000; Gu et al. 2002b). Wagner (2000) argued for decoupled evolution of genetic distance

**Table 2**  
**Possible Evolutionary Implications of Maximum-Likelihood Models of Gene Expression Evolution**

Class	Model	Possible Evolutionary Implications When Fit Is Good
Pure phylogenetic	Distance	Divergence in coding region predicts divergence in gene expression, would favor: neutralist model of evolution (22).
	Equal	Number of gene duplication events predicts variance in gene expression, a prediction of the subfunctionalization model (3).
	Free	Simpler models do not adequately describe evolution of gene expression, could mean: (a) sporadic changes in gene expression or (b) gene family does not function in process.
Nonphylogenetic	Distance	Genetic distances since last gene duplication predict change in expression, consistent with an initial coupling during evolution of expression and coding sequence (21).
	Equal	Closely related genes are no more likely than unrelated genes to have similar expression patterns, could mean: (a) Gene family does not function in process or (b) rapid rates of gene expression evolution.
	Free	Closely related genes are no more likely than unrelated genes to have similar expression patterns and expression change is sporadic, could mean: (a) Gene family does not function in process or (b) rapid rates of gene expression evolution.
Punctuated	Distance	After duplication, one daughter gene retains ancestral expression pattern, the other diverges in proportion to accumulated genetic distance of coding region, favors classical neofunctionalization (1).
	Equal	After duplication, one daughter gene retains ancestral expression pattern, the other diverges.
	Free	After duplication, one daughter gene retains ancestral function, the other changes in expression a variable amount.

and gene-expression divergence based on the lack of a significant correlation. However, Gu et al. (2002b) noted that the lack of correlation was driven by the inclusion of distantly related gene pairs. In fact, they found that a significant correlation does exist, when considering more closely related gene pairs. Gu et al. (2002b) also noted that synonymous distances may be a better proxy for evolutionary time, compared with overall genetic distance. Because in the current study we included genes separated by large genetic distances and we used an overall measure of genetic distance (not synonymous distance), it is not surprising that the present methods suggested little correlation between genetic distance and gene-expression divergence. Further underscoring this point is that in the current example, genetic distances since the most recent gene duplication (nonphylogenetic/distance model) events were better predictors of mRNA expression change than were overall genetic distances between two genes (pure phylogenetic/distance model), suggesting that older divergences obscured a possible correlation between genetic distance and gene-expression divergence. These results could also be further clarified by future studies using the current comparative approach, while utilizing synonymous distances.

Two well-known models of gene-duplication evolution deserve discussion in light of the current methods and results. First, the neofunctionalization hypothesis predicts that one copy of a duplicated gene may change function (Ohno 1970; Force et al. 1999). The punctuated models that we present here are similar to neofunctionalization; in both, one duplicate is assumed to diverge, whereas the other does not. However, a key difference does exist between our analyses and most previous studies. Our analyses utilize gene-expression data as a measure of functional divergence of genes, whereas many previous studies examined the coding region of genes, often using the ratio of nonsynonymous to synonymous rates as a different estimate of functional divergence.

Previous studies on substitution rates in coding regions have found conflicting results: some found that many duplicate gene pairs evolved at different rates, others found

little evidence of such rate differences. For example, duplicate frog genes lacked substitution rate differences when human genes were used as outgroups (Hughes and Hughes 1993; Hughes 1994), and duplicate genes of humans and rodents differed in rate in two of 49 cases in one study (Kondrashov et al. 2002). In contrast, Van de Peer et al. (2001) and Robinson-Rechavi and Laudet (2001) found significant rate differences in zebrafish duplicate genes, and Zhang, Gu, and Li (2003) found significantly different substitution rates in about 60% of recently duplicated human genes. These different results are caused by differences in the relationship of the outgroups to the duplicated genes and differences in data sets and statistical approaches.

In the current study, we found no support for our punctuated models, suggesting that yeast gene expression does not evolve by neofunctionalization. Our results are in contrast to those of Gu (2004), who found that 70% of three-member gene families showed significantly unequal rates of expression divergence after gene duplication.

Our lack of support for punctuated models may be influenced by two factors that deserve further discussion. First is the statistical approach that we used. The punctuated models summed likelihood values over all possible combinations of  $L_{n10}$  and  $L_{n01}$  from equations 3 and 4. We decided on this strategy because of ambiguity in the number of free parameters involved when using an alternative strategy of fixing the one of every pair of descendent branches to 0 that maximizes the overall likelihood (a maximum-likelihood analog of linear parsimony). By maximizing the likelihood in such a way, it is not clear if the act of choosing which branch is set to 0 increases the number of free parameters in the model, and if so, by how many. Summing over all possibilities avoids this ambiguity. However, the results may be different between the presented and alternative strategy. When using the alternative strategy, the punctuated models were among the best supported of all models if we did not add parameters for choosing which descendant branch to set to 0 (results not shown). An anonymous reviewer suggested that this problem could be

alleviated by using the Expectation-Maximization (E-M) algorithm, an endeavor that we leave for future work.

A second difficulty with rejecting the punctuated (neofunctionalization) models is that “extinction”—in this case, the loss of duplicated genes from a genome—will sometimes erase a punctuated evolutionary signal (Felsenstein 1988; Felsenstein 2004). For example, figure 4 shows a hypothetical phylogenetic tree of four genes whose branch lengths are punctuated, followed by trees resulting from extinction of different pairs of genes. Two of these trees maintain signature of a punctuated mode of evolution (fig. 4D and E), and two have lost the signature (see Felsenstein [2004] for a different example). Therefore, the interaction between loss of lineages and maintenance of a punctuated signal is complicated, and extinction may lead to the incorrect rejection of cases of truly punctuated evolution. An important consideration for how this point relates to the current study is that recently duplicated genes are less likely to have gone extinct compared with older duplications. Because we used a stringent criterion for grouping genes, most families probably contain recently duplicated genes, and so gene losses after duplication may be somewhat uncommon for the data set at hand.

In addition to neofunctionalization, the subfunctionalization hypothesis deserves discussion. Subfunctionalization predicts that ancestral genes have multiple functions that are lost in duplicated descendent genes in a complementary fashion (Force et al. 1999). Even though the current study on yeast may provide only a limited test of subfunctionalization, this model deserves discussion because it is widely cited, and it could be tested in future studies using methods similar to ours. The partitioning of gene function after duplication leads to the specific prediction that genes separated by more duplication events should have fewer functions in common. This is also a prediction of our pure phylogenetic/equal model, which was supported only in a few gene families based on AIC (fig. 3). Unfortunately, our current study on unicellular yeast precludes the examination of tissue-specific expression, which may be a primary driver of subfunctionalization. Nevertheless, our methods might have detected an analogous partitioning of expression among different timepoints. These results suggest that subfunctionalization may be more important in the evolution of multicellular organisms, because their multiple tissues may allow greater opportunity for partitioning of gene expression during evolution. This hypothesis could be addressed using methods similar to those presented here using comprehensive data on tissue-specific expression.

In summary, we maintain that different processes of gene-function evolution undoubtedly dominate in different taxa, in different gene families, and for different measures of function. Using functional genomic data, we can now begin large-scale empirical exploration of gene function, while focusing on gene expression. Our phylogenetic comparative approach to studying gene expression yielded results similar to previous sequence-based analyses: We report a lack of support for punctuated or neofunctionalization-like models. In addition, our comparative approach yielded results similar to previous studies of yeast gene-expression data that compared pairs of duplicated genes: When distantly

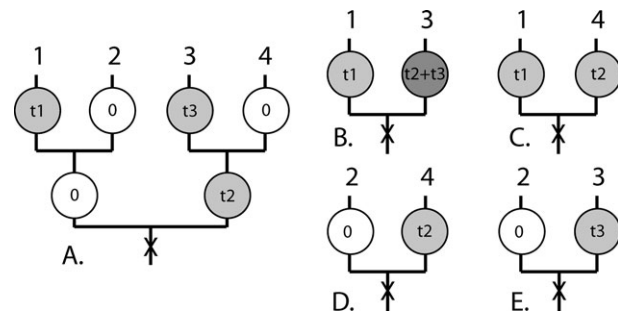


FIG. 4.—(A) A hypothetical case of punctuated evolution in a tree of four (1, 2, 3, 4) species (or genes), where  $t_x$  and 0 represent the amount of change in phenotype along a branch. (B–E) Extinction (loss) of different species (or genes) results in either maintenance (D and E) or loss (B and C) of the punctuated evolutionary signal. See Felsenstein (2004) for a different example illustrating the loss of punctuated signal due to lineage extinction.

related genes are included, genetic distance in the coding region of genes is a poor predictor of change in expression level (Wagner 2000), however for closely related genes, a correlation probably exists (Gu et al. 2002b).

### Future Directions

This study serves as an introduction to a general methodology for studying the evolution of gene expression, an important focus of modern evolutionary biology. The general approach we present can be applied to any species, group of species, or gene family and can test a wide variety of evolutionary hypotheses. Similar models also could be used to test for correlation between evolutionary changes in gene expression and phenotypic traits, allowing examination of hypotheses about the “genotype-phenotype map” (Fontana 2002; Murren 2002; Gompel and Carroll 2003; Sucena et al. 2003). Our methods could also be incorporated in a fully Bayesian statistical framework by placing prior distributions on the parameters in equations 1 and 2 (with methods similar to Huelsenbeck and Rannala [2003]) and estimating phylogenetic trees using a Bayesian approach (e.g., Rannala and Yang 1996; Yang and Rannala 1997). The methods would also benefit from simulation studies to test their power. A final important extension is to develop similar methods that utilize the spatial information of in situ hybridization or reporter gene assays, which are also now becoming available on genome-wide scales (Harafuji, Keys, and Levine 2002; Satou et al. 2002; Wada et al. 2003).

### Acknowledgments

We thank J. Felsenstein, M. Hahn, P. Lewis, A. Mooers, M. Holder, M. Rockman, D. Swofford, and K. Thornton for suggestions. This study was supported by NIH grant GM66104 to W.-H.L. and an NIH NRSA fellowship to T.H.O.

### Literature Cited

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pp. 267–281. *Proceeding of*

- the 2nd International Symposium on Information Theory, Supplement. Problems of control and information theory.
- Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. *Science* **165**:349–357.
- Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. 1998. The transcriptional program of sporulation in budding yeast. *Science* **282**:699–705.
- DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**:680–686.
- Diaz-Uriarte, R., and T. Garland. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. *Syst. Biol.* **45**:27–47.
- Edwards, A. W. F. 1992. Likelihood. Johns Hopkins University Press, Baltimore.
- Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**:471–492.
- . 1981a. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- . 1981b. Evolutionary trees from gene-frequencies and quantitative characters—finding maximum likelihood estimates. *Evolution* **35**:1229–1242.
- . 1988. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.* **19**:445–471.
- . 1985. Phylogenies and the comparative method. *Am. Nat.* **125**:1–15.
- . 1995. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.
- . 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.
- Ferea, T. L., D. Botstein, P. O. Brown, and R. F. Rosenzweig. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. USA* **96**:9721–9726.
- Fontana, W. 2002. Modelling ‘evo-devo’ with RNA. *Bioessays* **24**:1164–1177.
- Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531–1545.
- Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**:4241–4257.
- Gompel, N., and S. B. Carroll. 2003. Genetic mechanisms and constraints governing the evolution of correlated traits in drosophilid flies. *Nature* **424**:931–935.
- Gu, X. 2004. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics* **167**:531–542.
- Gu, Z., A. Cavalcanti, F. C. Chen, P. Bouman, and W. H. Li. 2002a. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* **19**:256–262.
- Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li. 2002b. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* **18**:609–613.
- Harafuji, N., D. N. Keys, and M. Levine. 2002. Genome-wide identification of tissue-specific enhancers in the *Ciona* tadpole. *Proc. Natl. Acad. Sci. USA* **99**:6802–6805.
- Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTAL V: Improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**:189–191.
- Huelsenbeck, J. P., and B. Rannala. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evol. Int. J. Org. Evol.* **57**:1237–1247.
- Hughes, A. L. 1994. The evolution of functionally novel proteins after gene duplication. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **256**:119–124.
- Hughes, M. K., and A. L. Hughes. 1993. Evolution of duplicated genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10**:1360–1369.
- King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**:107–116.
- Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Selection in the evolution of gene duplications. *Genome Biol.* **3**:RESEARCH0008.
- Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**:913–925.
- Lyons, T. J., A. P. Gasch, L. A. Gaither, D. Botstein, P. O. Brown, and D. J. Eide. 2000. Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proc. Natl. Acad. Sci. USA* **97**:7957–7962.
- Makova, K. D., and W. H. Li. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* **13**:1638–1645.
- Martins, E. P., and T. Garland Jr. 1991. Phylogenetic analysis of the correlated evolution of continuous characters: a simulation study. *Evolution* **45**:534–557.
- Mayr, E., and W. B. Provine. 1980. The evolutionary synthesis: perspectives on the unification of biology. Harvard University Press, Cambridge, Mass.
- Mooers, A. Ø., and D. Schluter. 1998. Fitting macroevolutionary models to phylogenies: an example using vertebrate body sizes. *Contrib. Zool.* **68**:3–18.
- Mooers, A. Ø., S. M. Vamossi, and D. Schluter. 1999. Using phylogenies to test macroevolutionary hypotheses of trait evolution in Cranes (Gruinae). *Am. Nat.* **154**:249–259.
- Murren, C. J. 2002. Phenotypic integration in plants. *Plant Spec. Biol.* **17**:89–99.
- Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, New York.
- Pagel, M. D. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B Biol. Sci.* **255**:37–45.
- . 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48**:612–622.
- Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**:817–818.
- Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**:304–311.
- Robinson-Rechavi, M., and V. Laudet. 2001. Evolutionary rates of duplicate genes in fish and mammals. *Mol. Biol. Evol.* **18**:681–683.
- Rost, B. 1999. Twilight zone of protein sequence alignments. *Prot. Eng.* **12**:85–94.
- Royall, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman and Hall, London.
- Satou, Y., L. Yamada, Y. Mochizuki et al (14 co-authors). 2002. A cDNA resource from the basal chordate *Ciona intestinalis*. *Genesis* **33**:153–154.
- Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**:3273–3297.
- Sucena, E., I. Delon, I. Jones, F. Payre, and D. L. Stern. 2003. Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism. *Nature* **424**:935–938.

- Swofford, D. L. 1999. PAUP\*: phylogenetic analysis using parsimony (\*and other methods). Sinauer Associates, Sunderland, Mass.
- Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**:512–526.
- Van de Peer, Y., J. S. Taylor, I. Braasch, and A. Meyer. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J. Mol. Evol.* **53**:436–446.
- Wada, S., M. Tokuoka, E. Shoguchi et al. (13 co-authors). 2003. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. II. Genes for homeobox transcription factors. *Dev. Genes Evol.* **213**:222–234.
- Wagner, A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci. USA* **97**:6579–6584.
- Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**:1377–1419.
- Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14**:717–724.
- Zhang, P., Z. Gu, and W. H. Li. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* **4**:R56.

Arndt von Haeseler, Associate Editor

Accepted August 30, 2004